УДК 57.087.1

РАЗРАБОТКА АЛГОРИТМОВ АНАЛИЗА БИОЧИПОВ ДНК С УЧЕТОМ ВЕСОВЫХ ФАКТОРОВ КАЧЕСТВА СПОТОВ

И.В. КЛИМУК, А.С. СВИДРИЦКИЙ, канд. физ.-мат. наук, доц. Н.Н. ЯЦКОВ
(Белорусский государственный университет, Минск)

В работе предложены три модификации алгоритма *k*-ближайших соседей для анализа биочипов ДНК с учетом весовых факторов качества спотов. Представлены результаты сравнения алгоритмов *k*-ближайших соседей без учета и с учетом параметра качества спотов на примере смоделированных данных.

Биочипы или микрочипы ДНК – это микроматрицы с нанесенными на них образцами биологического вещества [1]. Особенностью микрочипов является возможность одновременно исследовать экспрессии множества генов [2]. С помощью технологии биочипов ДНК можно за короткое время обнаружить различные онкологические заболевания, туберкулез и др. [3]. Для более эффективного использования микрочипов требуется постоянное улучшение алгоритмов их анализа. Однако часто обработка и анализ биочипа затруднены вследствие низкого качества данных и высокого уровня экспериментального шума. Повысить эффективность алгоритмов анализа биочипов, а именно классификации экспрессии генов, можно, если учесть параметр качества изображения каждого спота биочипа [4, 5]. Возможность вычисления параметров или факторов качества спотов предоставляют некоторые программные пакеты цифровой обработки изображений биочипов, например, MAIA [6].

Цель работы: реализация и исследование алгоритма *k*-ближайших соседей для анализа биочипов ДНК с учетом весовых факторов качества спота.

В работе приведены результаты сравнения алгоритмов *k*-ближайших соседей без учета и с учетом параметра качества спотов.

ИМИТАЦИОННАЯ МОДЕЛЬ БИОЧИПА

В работе использовалась модель двухканального биочипа ДНК. За основу взята модель [7]. Она не требует знания физических процессов, происходящих во время экспериментов с биочипами. Алгоритм моделирования:

Шаг 1. Задание параметров модели: *N* – число генов, *m* – число репликантов, *p* – доля невыраженных генов в выборке.

Шаг 2. Создание заполнение матрицы *M* размером *N*m*: первые *N*p* строк – значением 0, следующие *N*(1-p)/2* строк – значением -1, последние *N*(1-p)/2* строк – значением 1.

Шаг 3. Генерация вектора параметров качества спотов *Q* размером *N* с использованием бета-распределения с параметрами a = 2.5, b = 3.5. Данное распределение позволяет добиться генерации значений параметров качества спотов, наиболее близко имитирующих значения реальных экспериментов [5].

Шаг 4. Добавление нормального шума к данным по формуле (1):

$$M_i = M_i + rnorm * (1 - Q_i), \quad (1)$$

где *rnorm* – реализация стандартной нормальной случайной величины.

Строки полученной матрицы *M* – объекты, которые требуется классифицировать, вектор *Q* – вектор параметров качества каждого объекта.

МЕТОД *k*-БЛИЖАЙШИХ СОСЕДЕЙ И ЕГО МОДИФИКАЦИИ

В общем случае задача классификации представляет из себя разбиение конечного множества объектов (тестируемая выборка) на классы, имея некоторое значительно меньшее множество с уже известными метками классов (обучающая выборка). В случае с биочипами ДНК на этапе количественного анализа нужно классифицировать гены как невыраженные, выраженные или подавленные. В идеальном случае каждый ген имел бы значение относительной экспрессии 0, 1 и -1 соответственно. Однако в связи с высоким экспериментальным шумом задача классификации усложняется.

В данной работе для исследования выбран и программно реализован метод *k*-ближайших соседей (сокращенно kNN). Его преимущества: простота реализации, достаточно хорошие результаты при работе с данными различного типа, возможность модифицировать и подстроить алгоритм под конкретную задачу. Недостатки: большая трудоемкость в вычислениях, неопределенность выбора числа *k.*

Разработаны и программно реализованы 3 модификации метода с учетом весовых факторов качества спотов.

1) kNN с учетом параметра качества спота при расчете дистанции. Расстояния между объектами обучающей и тестируемой выборки считаются по формуле (2):

$$d_{ij}^{w} = {}^{d_{ij}}\!/_{q_j} \quad , \tag{2}$$

где q_j – параметр качества *j-го* объекта обучающей выборки.

2) kNN с учетом параметра качества спота при голосовании. Параметр качества учитывается при назначении метки класса объекту тестируемой выборки. Класс-победитель выбирается по максимальной сумме их параметров качества (3):

$$Class(n_i) = arg \max_{c \in C} \sum_{j=1}^{l} [c_j = c]\, q_j \tag{3}$$

3) kNN с учетом параметра качества спота при расчете дистанций и при голосовании. Алгоритм включает оба изменения, описанные в п. 1) и 2)

СРАВНЕНИТЕЛЬНЫЙ АНАЛИЗ РАБОТЫ АЛГОРИТМОВ

Эффективность работы каждого алгоритма считалась как процент ошибки при классификации всех генов, невыраженных и выраженных генов (представляющих наибольший интерес при анализе биочипов).

Выполнено сравнение эффективности алгоритмов в зависимости от различных значений параметров обучающей и тестируемой выборки, а именно:

1) N_L – размер обучающей выборки менялся от 100 до 2000

2) p_L – относительная доля невыраженных генов в обучающей выборке менялось от 0.05 до 0.95

3) <Q> и <Q_L > – средние значения параметров качества тестируемой и обучающей выборки соответственно. Оба менялись от 0.1 до 0.9

Зависимость ошибки классификации от параметров модели для разработанных алгоритмов представлены на рисунках 1, 2, 3, 4. Видно, что модифицированные алгоритмы имеют меньшую ошибку классификации, чем классический алгоритм.

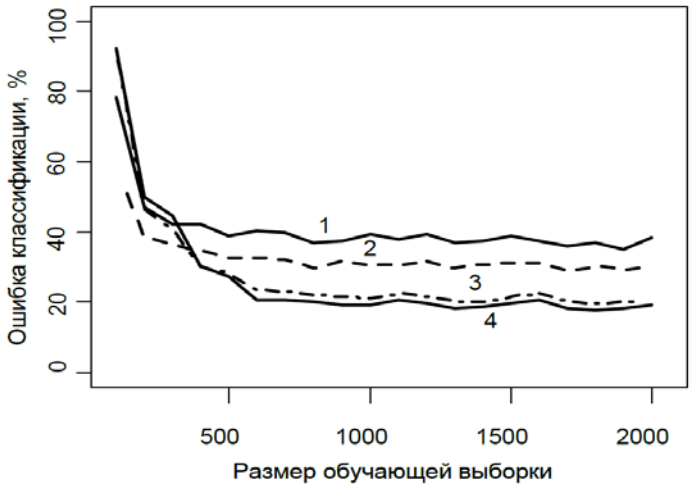

Рисунок 1. – Зависимость ошибки классификации выраженных генов от размера обучающей выборки: 1 – kNN, 2 – kNN с измененным голосованием, 3 – kNN с пересчетом расстояний, 4 – kNN с измененным голосованием и пересчетом расстояний

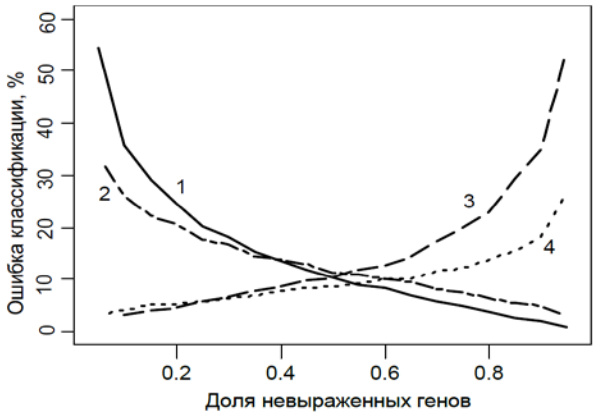

Рисунок 2. – Зависимость ошибки классификации от доли невыраженных генов в обучающей выборке: 1 – невыраженных генов для kNN, 2 – невыраженных генов для kNN с измененным голосованием и пересчетом расстояний, 3 – выраженных генов для kNN, 4 – выраженных генов для kNN с измененным голосованием и пересчетом расстояний

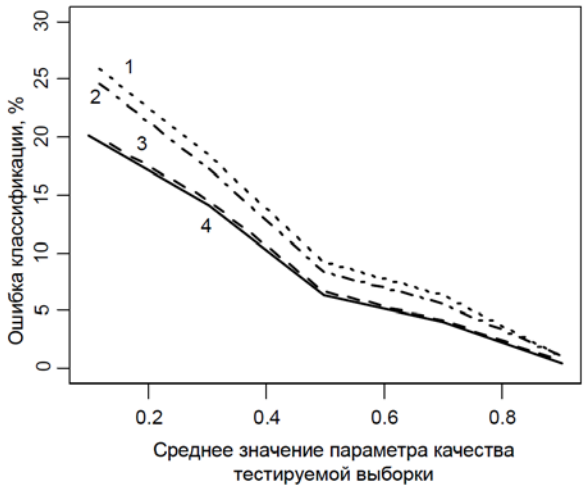

Рисунок 3. – Зависимость ошибки классификации выраженных генов от среднего качества тестируемой выборки: 1 – kNN, 2 – kNN с измененным голосованием, 3 – kNN с пересчетом расстояний, 4 – kNN с измененным голосованием и пересчетом расстояний

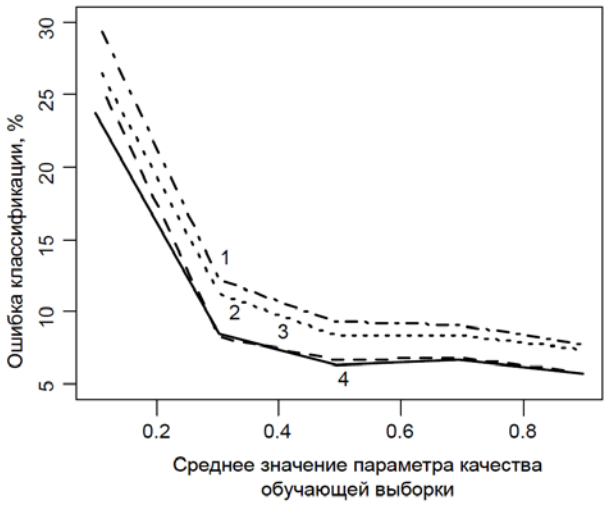

Рисунок 4. – Зависимость ошибки классификации выраженных генов от среднего качества обучающей выборки: 1 – kNN, 2 – kNN с измененным голосованием, 3 – kNN с пересчетом расстояний, 4 – kNN с измененным голосованием и пересчетом расстояний

ВЫВОДЫ. Разработаны и реализованы метод *k*-ближайших соседей и три его модификации с учетом параметра качества спотов. Сравнительный анализ эффективности работы алгоритмов позволяет сделать выводы:

1) при размере обучающей выборки 500 и более ошибка классификации для всех алгоритмов не меняется;

2) чем хуже качество данных (как обучающей выборки, так и тестируемой), тем лучше с задачей классификации справляются модифицированные алгоритмы;

3) наилучший алгоритм - *k*-ближайших соседей с учетом параметра качества для расчета расстояний и при голосовании: ошибка классификации на 7% ниже, чем у классического алгоритма.

Литература

1. Microarray and whole-exome sequencing analysis of familial Beh¸cet's disease patients / D. Okuzaki [et al.]. Режим доступа: URL://www.ncbi.nlm.nih.gov/pmc/articles/PMC4726226/.
2. Мирзобеков, А.Д. Биочипы в биологии и медицине 21го века / А.Д. Мирзобеков // Вестник Российской Академии Наук. – 2003. – Т. 73., № 5. – С. 412.
3. Zou, J. Analysis of microarray-identified genes and microRNAs associated with drug resistance in ovarian cancer/ J. Zou, F. Yin, Q. Wang // International Journal of Clinical and Experimental Pathology. – 2015.
4. Novikov, E. An algorithm for automatic evaluation of the spot quality in two-color DNA microarray experiments / E. Novikov, E. Barillot // BMC Bioinformatics. – 2005.
5. Advanced spot quality analysis in two-colour microarray experiments / M. Yatskou [et al.] // BMC Research Notes. – 2008.
6. Novikov, E. Software package for automatic microarray image analysis (MAIA) / E. Novikov, E. Barillot // BMC Bioinformatics. – 2007. – V. 5. – P. 639.
7. Demb´el´e, D. A Flexible Microarray Data Simulation Model / D. Demb´el´e // Microarrays. – 2013.